PHYSICAL REVIEW E **68**, 021913 (2003)

Multifractal and correlation analyses of protein sequences from complete genomes

Zu-Guo Yu,[1,2,*] Vo Anh,[1] and Ka-Sing Lau[3]

[1]*Program in Statistics and Operations Research, Queensland University of Technology, GPO Box 2434, Brisbane Q4001, Australia*

[2]*Department of Mathematics, Xiangtan University, Hunan 411105, China*

[3]*Department of Mathematics, Chinese University of Hong Kong, Shatin, Hong Kong, China*

(Received 21 March 2003; published 22 August 2003)

A measure representation of protein sequences similar to the measure representation of DNA sequences proposed in our previous paper [Yu *et al.*, Phys. Rev. E **64**, 031903 (2001)] and another induced measure are introduced. Multifractal analysis is then performed on these two kinds of measures of a large number of protein sequences derived from corresponding complete genomes. From the values of the D_q (generalized dimensions) spectra and related C_q (analogous specific heat) curves, it is concluded that these protein sequences are not completely random sequences. For substrings with length $K=5$, the D_q spectra of all organisms studied are multifractal-like and sufficiently smooth for the C_q curves to be meaningful. The C_q curves of all bacteria resemble a classical phase transition at a critical point. But the ''analogous'' phase transitions of higher organisms studied exhibit the shape of double-peaked specific heat function. But for the classification problem, the multifractal property is not sufficient. When the measure representations of protein sequences from complete genomes are considered as time series, a method based on correlation analysis after removing some memory from the time series is proposed to construct a phylogenetic tree. This construction is shown to be reasonably satisfactory.

DOI: 10.1103/PhysRevE.68.021913 PACS number(s): 87.14.Gg

I. INTRODUCTION

Since the sequencing of the first complete genome of the free-living bacterium *Mycoplasma genitalium* in 1995 [1], more and more complete genomes have been deposited in public databases such as Genbank [34]. Complete genomes provide essential information for understanding gene functions and evolution. To be able to determine the patterns of DNA and protein sequences is very useful for studying many important biological problems such as identifying new genes and establishing the phylogenetic relationship among organisms.

A DNA sequence is formed by four different nucleotides, namely, adenine (a), cytosine (c), guanine (g), and thymine (t). A protein sequence is formed by 20 different kinds of amino acids, namely, alanine (A), arginine (R), asparagine (N), aspartic acid (D), cysteine (C), glutamic acid (E), glutamine (Q), glycine (G), histidine (H), isoleucine (I), leucine (L), lysine (K), methionine (M), phenylalanine (F), proline (P), serine (S), threonine (T), tryptophan (W), tyrosine (Y), and valine (V) (Ref. [2], p. 109). The protein sequences from complete genomes are translated from their coding sequences (DNA) through the genetic code (Ref. [2], p. 122).

A useful result is the establishment of long memory in DNA sequences [3–6]. Li and Kanero [3] found that the spectral density of a DNA sequence containing mostly introns shows $1/f^{\beta}$ behavior, which indicates the presence of long-range correlation when $0<\beta<1$. The correlation properties of coding and noncoding DNA sequences were also studied by Peng *et al.* [4] in their fractal landscape or DNA walk model. Peng *et al.* [4] discovered that there exists long-range correlation in noncoding DNA sequences while the coding sequences correspond to a regular random walk. By undertaking a more detailed analysis, Chatzidimitriou-Dreismann and Larharmmar [5] concluded that both coding and noncoding sequences exhibit long-range correlation. A subsequent work by Prabhu and Claverie [6] also corroborated these results. From a different angle, fractal analysis is a relatively new analytical technique that has proved useful in revealing complex patterns in natural phenomena. Berthelsen *et al.* [7] considered the global fractal dimension of human DNA sequences treated as pseudorandom walks. Vieira [8] carried out a low-frequency analysis of the complete DNA of 13 microbial genomes and showed that their fractal behavior does not always prevail through the entire chain and the autocorrelation functions have a rich variety of behaviors including the presence of antipersistence.

Although statistical analyses performed directly on DNA sequences have yielded some success, there has been some indication that this method is not powerful enough to amplify the difference between a DNA sequence and a random sequence as well as to distinguish DNA sequences themselves in more details [9]. One needs more powerful global and visual methods. For this purpose, Hao *et al.* [9] proposed a visualization method based on counting and coarse graining the frequency of appearance of substrings with a given length. They called it the *portrait* of an organism. They found that there exist some fractal patterns in the portraits which are induced by avoiding and underrepresented strings. The fractal dimension of the limit set of portraits was also discussed [10,11]. There are other graphical methods of sequence patterns, such as the chaos game representation [12,13].

Multifractal analysis is a useful way to characterize the spatial heterogeneity of both theoretical and experimental

*Corresponding author. Email address: yuzg@hotmail.com or z.yu@qut.edu.au

fractal patterns [14]. Yu *et al.* [15] introduced a representation of a DNA sequence by a probability measure of K strings derived from the sequence. This probability measure is in fact the histogram of the events formed by all the K strings in a dictionary ordering. It was found [15] that these probability measures display a distinct multifractal behavior characterized by their generalized Rényi dimensions (instead of a single fractal dimension as in the case of self-similar processes). Furthermore, the corresponding C_q curves (defined in Ref. [16]) of these generalized dimensions of all bacteria resemble classical phase transition at a critical point, while the ''analogous'' phase transitions (defined in Ref. [16]) of chromosomes of nonbacteria exhibit the shape of double-peaked specific heat function. These patterns led to a meaningful grouping of archaebacteria, eubacteria, and eukaryote. Anh *et al.* [17] took a further step in providing a theory to characterize the multifractality of the probability measures of complete genomes. In particular, the resulting parametric models fit extremely well the D_q curves of the generalized dimensions and the corresponding K_q curves of the above probability measures of the complete genomes. Based on the measure representation of DNA sequence and the technique of multifractal analysis in Ref. [15], Anh *et al.* [18] discussed the problem of recognition of an organism from fragments of its complete genome.

Works have been done to study the phylogenetic relationship based on correlation analyses of the K strings of complete genomes [19] and protein sequences from complete genomes [20,21]. Qi *et al.* [20] pointed out that a phylogenetic tree based on the protein sequences from complete genomes is more precise than a tree based on the complete genomes (DNA) themselves, and removing the random background from the probabilities of K strings of protein sequences can improve a phylogenetic tree from the biological point of view.

In this direction, we introduce in this paper the notion of measure representation of protein sequences similar to that of DNA sequences introduced in Ref. [15]. We then perform multifractal analyses on this kind of measure representation of protein sequences. We also construct a different measure by subtracting some memory from the original measure. Then multifractal analyses are performed on these new measures, and a phylogenetic tree is constructed based on their correlation analyses.

II. MEASURE REPRESENTATION

Each coding sequence in the complete genome of an organism can be translated into a protein sequence using the genetic code (Ref. [2], p. 122). Then we can link all translated protein sequences from a complete genome to form a long protein sequence according to the order of the coding sequences in the complete genome. In this way, we obtain a linked protein sequence for each organism. In this paper we only consider this kind of linked protein sequences and view them as symbolic sequences.

We call any string made of K letters from the alphabet $\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$ which corresponds to 20 kinds of amino acids a K string. For a given K, there are in total 20^K different K strings. In order to count the number of each kind of K strings in a given protein sequence, 20^K counters are needed. We divide the interval $[0,1[$ into 20^K disjoint subintervals, and use each subinterval to represent a counter. Letting $s=s_1\cdots s_K$, $s_i\in\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$, $i=1,\ldots,K$, be a substring with length K, we define

$$x_l(s)=\sum_{i=1}^{K}\frac{x_i}{20^i}, \tag{1}$$

where

$$x_i=\begin{cases} 0, & \text{if } s_i=A,\\ 1, & \text{if } s_i=C,\\ 2, & \text{if } s_i=D,\\ 3, & \text{if } s_i=E,\\ 4, & \text{if } s_i=F,\\ 5, & \text{if } s_i=G,\\ 6, & \text{if } s_i=H,\\ 7, & \text{if } s_i=I,\\ 8, & \text{if } s_i=K,\\ 9, & \text{if } s_i=L,\\ 10, & \text{if } s_i=M,\\ 11, & \text{if } s_i=N,\\ 12, & \text{if } s_i=P,\\ 13, & \text{if } s_i=Q,\\ 14, & \text{if } s_i=R,\\ 15, & \text{if } s_i=S,\\ 16, & \text{if } s_i=T,\\ 17, & \text{if } s_i=V,\\ 18, & \text{if } s_i=W,\\ 19, & \text{if } s_i=Y, \end{cases} \tag{2}$$

and

$$x_r(s)=x_l(s)+\frac{1}{20^K}. \tag{3}$$

We then use the subinterval $[x_l(s),x_r(s)[$ to represent substring s. Let $N_K(s)$ be the number of times that substring s with length K appears in the linked protein sequence and $N_K(total)$ the total times of all substrings with length K appear in the linked protein sequence [we use an open reading frame and slide one position each time to count the times; $N_K(s)$ may be zero]. We define

$$F_K(s)=N_K(s)/N_K(total) \tag{4}$$

to be the frequency of substring s. It follows that $\Sigma_{\{s\}}F_K(s)=1$. We can now define a measure μ_K on $[0,1[$ by $d\mu_K(x)=Y_K(x)dx$, where

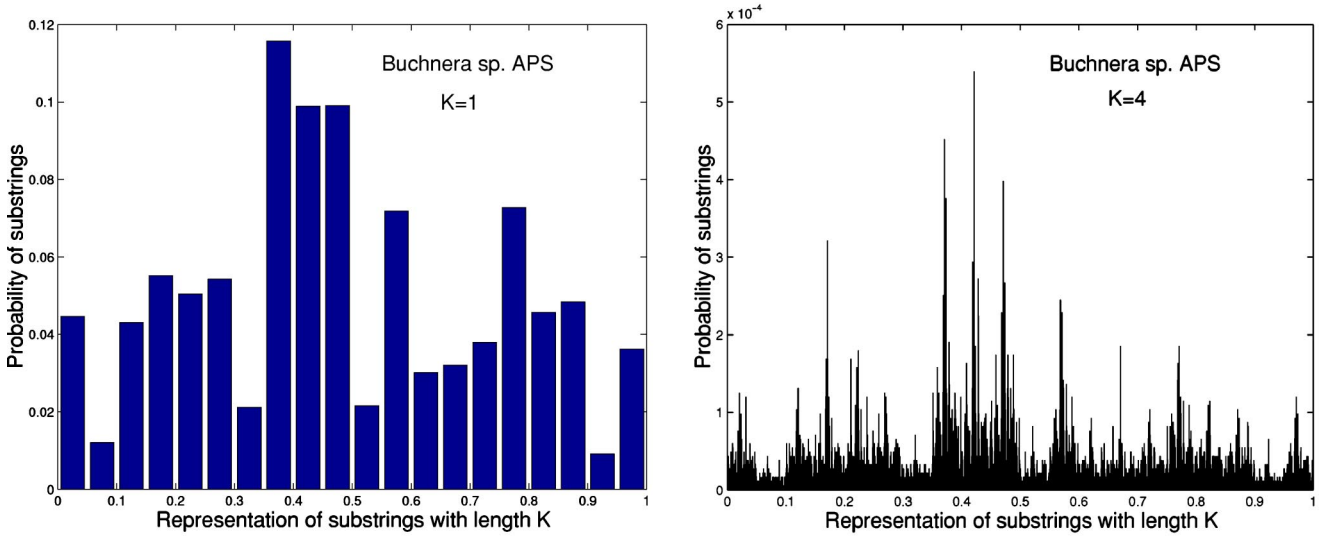

FIG. 1. Histograms of substrings with lengths $K=1$ and 4 of protein sequence from complete genome of *Buchnera sp. APS*

$$Y_K(x)=20^K F_K(s), \quad \text{when} \quad x\in[x_l(s),x_r(s)[. \tag{5}$$

It is easy to see that $\int_0^1 d\mu_K(x)=1$ and $\mu_K([x_l(s),x_r(s)[)=F_K(s)$. We call μ_K the *measure representation* of the linked protein sequence of the organism corresponding to the given K. As an example, the histogram of substrings in the linked protein sequence of *Buchnera sp. APS* for $K=1$ and 4 are given in Fig. 1.

For simplicity of notation, the index K is dropped in $F_K(s)$, etc., from now on, where its meaning is clear. We can order all the $F(s)$ according to the increasing order of $x_l(s)$. We then obtain a sequence of real numbers consisting of 20^K elements which we denote as $F(t), t=1,\ldots,20^K$.

Remark 1. As in Ref. [15], the ordering of 20 letters in Eq. (2) follows the natural dictionary ordering of K strings in the one-dimensional space. A different ordering of 20 letters would change the correlation structure of the measure. However, by its construction, different orderings of 20 letters in Eq. (2) give almost the same multifractal spectrum and the D_q curve, which will be defined in the following section, when the absolute value of q is relatively small (In Ref. [15] we have the same property for the measure representation of DNA sequence). We shown in Fig. 2 the D_q curves for four different orderings to support this statement. Hence, our results based on multifractal analysis are considered independent of the ordering. In a comparison of different organisms using this measure representation, once the ordering is given, it is fixed for all organisms.

If s' is one of the 20 letters, we denote by $P(s')$ the frequency of letter s' in the linked protein sequence. Then for any K substring $s=s_1\cdots s_K$, $s_i\in\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$, $i=1,\ldots,K$, we define

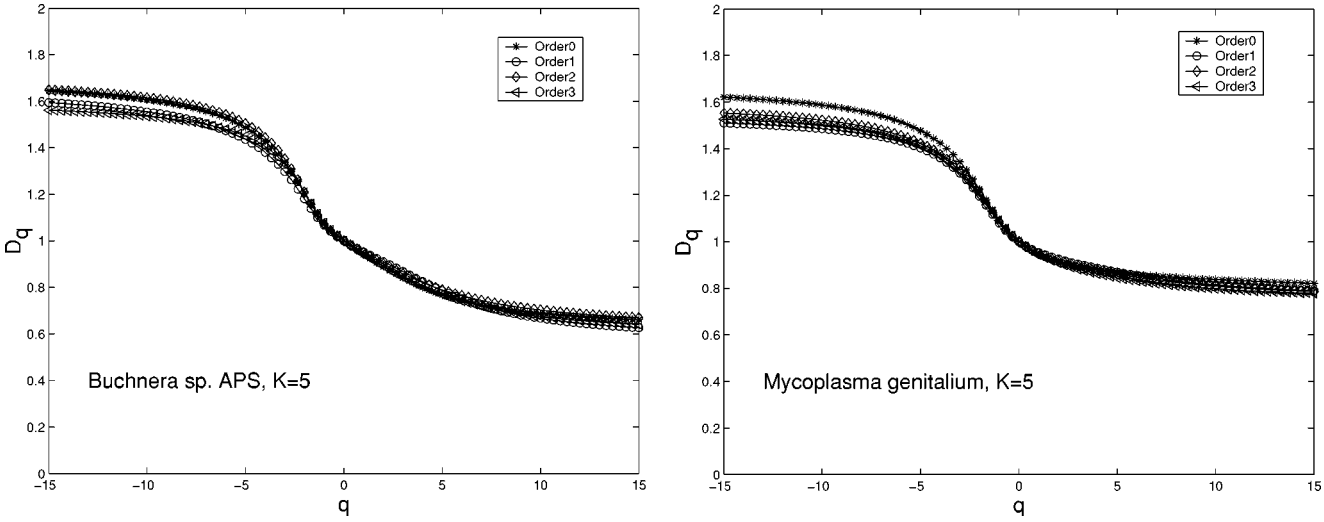

FIG. 2. The D_q curves based on four different orders of the 20 kinds of amino acids in Eq. (2) for measure representation of the linked protein sequences of *Buchnera sp. APS* (left) and *Mycoplasma genitalium* (right). Order0 is the dictionary order; Order1 is $\{L,A,M,C,N,D,P,E,Q,F,R,G,S,H,T,I,V,K,Y,W\}$; Order2 is $\{C,L,D,M,E,N,F,P,G,Q,H,R,I,S,K,T,V,W,Y,A\}$; and Order3 is $\{Y,C,W,A,L,D,M,E,N,F,P,G,Q,H,R,I,S,K,V,T\}$.

$$F'(s)=P(s_1)P(s_2)\cdots P(s_K).$$

We next define

$$F^d(s)=F(s)-F'(s) \tag{6}$$

and denote by $F^{ad}(s)$ the absolute value of $F^d(s)$. For all 20^K different K strings, we can also order the $F^d(s)$ sequence and $F^{ad}(s)$ sequence according to the dictionary order of s.

From the point of view of Ref. [20], we need to subtract the random background from the sequence $\{F(s)\}$ in order to get a more satisfactory evolutionary tree. Qi *et al.* used a Markov model to do this. Here we use the frequencies of the 20 kinds of amino acids appearing in the linked protein sequence. By the nature of its generation, this probability measure behaves as a multiplicative cascade and displays long memory. Hence, subtracting out the fractal background $F'(s)$ as described above has the effect of reducing long memory in the measure representation.

Based on the sequence $\{F^{ad}(s)\}$, we obtain a different measure via a similar way described above (see also Ref. [22]) after normalization. We denote this measure by μ'.

III. MULTIFRACTAL ANALYSIS AND CORRELATION ANALYSIS

Common numerical implementation of multifractal analysis is based on the *fixed-size box-counting algorithms* [23]. In the one-dimensional case, for a given measure μ with support $E\subset\mathbf{R}$, we consider the *partition sum*

$$Z_\epsilon(q)=\sum_{\mu(B)\neq 0}[\mu(B)]^q, \tag{7}$$

$q\in\mathbf{R}$, where the sum runs over all different nonempty boxes B of a given side ϵ in a grid covering of the support E, that is,

$$B=[k\epsilon,(k+1)\epsilon[. \tag{8}$$

The exponent $\tau(q)$ is defined by

$$\tau(q)=\lim_{\epsilon\to 0}\frac{\ln Z_\epsilon(q)}{\ln\epsilon} \tag{9}$$

and the generalized fractal dimensions of the measure are defined as

$$D_q=\tau(q)/(q-1), \quad \text{for} \quad q\neq 1 \tag{10}$$

and

$$D_q=\lim_{\epsilon\to 0}\frac{Z_{1,\epsilon}}{\ln\epsilon}, \quad \text{for} \quad q=1, \tag{11}$$

where $Z_{1,\epsilon}=\Sigma_{\mu(B)\neq 0}\mu(B)\ln\mu(B)$. The generalized fractal dimensions are numerically estimated through a linear regression of

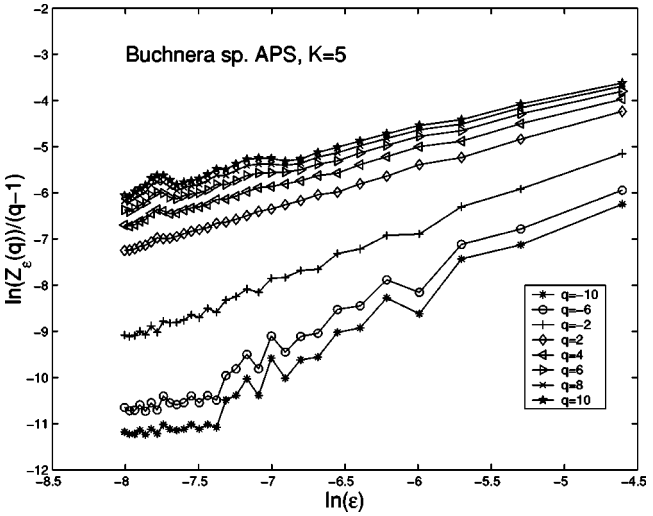

FIG. 3. The linear slopes in the D_q spectra.

$$\frac{1}{q-1}\ln Z_\epsilon(q)$$

against $\ln\epsilon$ for $q\neq 1$, and similarly through a linear regression of $Z_{1,\epsilon}$ against $\ln\epsilon$ for $q=1$. For example, we show how to obtain the D_q spectrum using the slope of the linear regression in Fig. 3. D_1 is the *information dimension* and D_2 is the *correlation dimension* of the measure. The D_q of the positive values of q give relevance to the regions where the measure is large, i.e., to the K strings with high probability. The D_q of the negative values of q are associated with the structure and properties of the most rarefied regions of the measure.

Figure 3 shows that the linear fitting becomes relatively worse when the absolute value of q increases. In order to overcome the finite-size effects (due to the small size of a single protein) and to attain statistical convergence, all the protein sequences, translated from coding sequences in the complete genome, are linked together into a long sequence of proteins which we called a linked protein sequence. For such an extended sequence, the size is sufficiently long for the asymptotic results of multifractal analysis to hold or be approximately correct. Second, the values of D_q used in this study are those corresponding to q with smaller absolute values, and as a result the estimation is fairly accurate.

Some sets of physical interest have a nonanalytic dependence of D_q on q. Moreover, this phenomenon has a direct analogy to the phenomenon of phase transitions in condensed-matter physics [24]. The existence and type of phase transitions might turn out to be a worthwhile characterization of universality classes for the structures [25]. The concept of phase transition in multifractal spectra was introduced in the study of logistic maps, Julia sets, and other simple systems. Evidence of phase transition was found in the multifractal spectrum of diffusion-limited aggregation [26]. By following the thermodynamic formulation of multifractal measures, Canessa [16] derived an expression for the analogous specific heat as

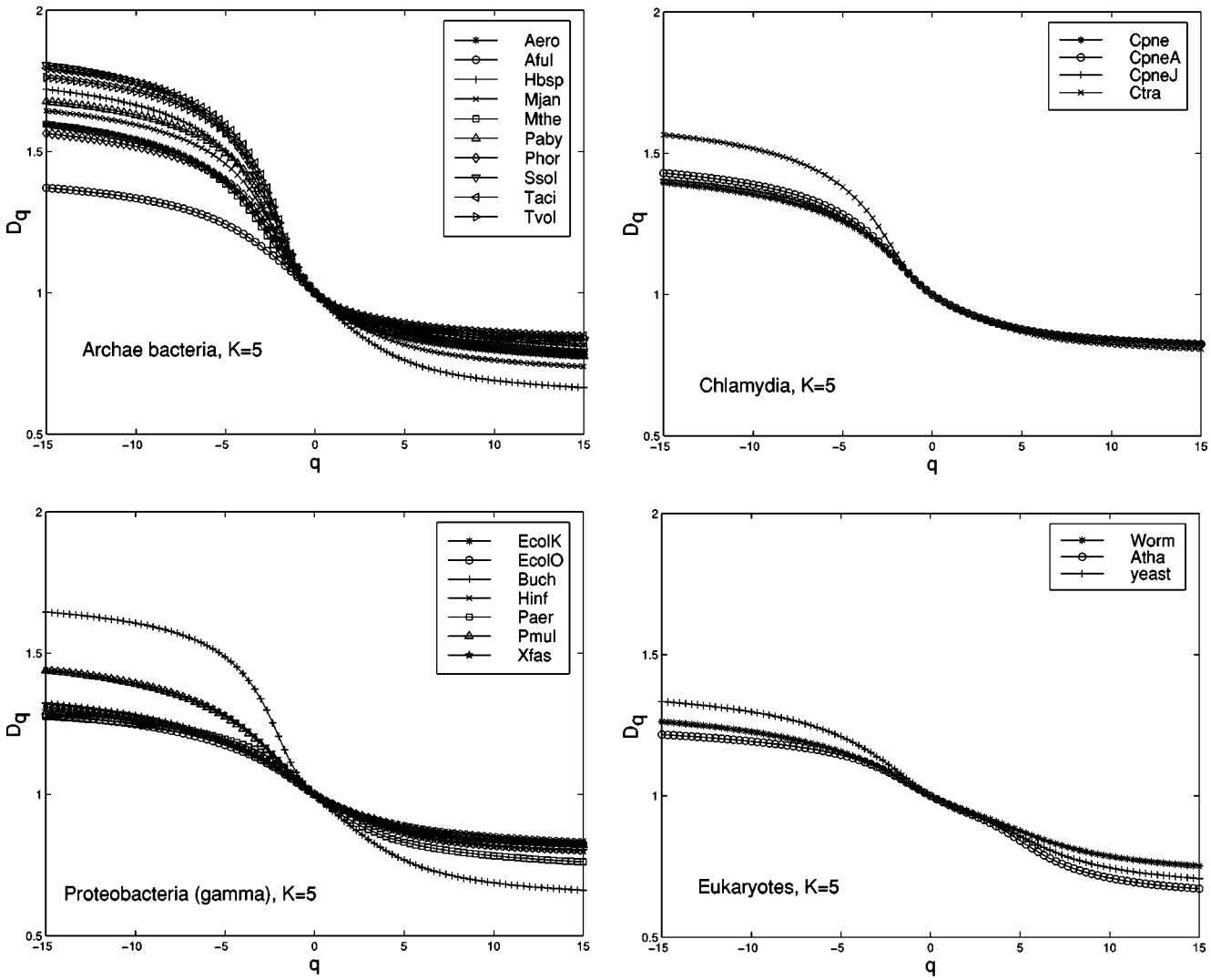

FIG. 4. Dimension spectra of measure representation μ of protein sequences of some organisms.

$$C_q \equiv -\frac{\partial^2 \tau(q)}{\partial q^2} \approx 2\tau(q) - \tau(q+1) - \tau(q-1). \quad (12)$$

He showed that the form of C_q resembles a classical phase transition at a critical point for financial time series. In the following section, we discuss the property of C_q for measures μ and μ' defined in Sec. II.

For two random variables $\mathbf{X}$ and $\mathbf{Y}$ with samples $X(1), X(2), \ldots, X(N)$ and $Y(1), Y(2), \ldots, Y(N)$, respectively, let

$$\langle \mathbf{X} \rangle = \frac{1}{N} \sum_{i=1}^{N} X(i), \quad \langle \mathbf{Y} \rangle = \frac{1}{N} \sum_{i=1}^{N} Y(i),$$

$$\delta(\mathbf{X}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} [X(i) - \langle \mathbf{X} \rangle]^2},$$

$$\delta(\mathbf{Y}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} [Y(i) - \langle \mathbf{Y} \rangle]^2}.$$

Then, the sample covariance of $\mathbf{X}$ and $\mathbf{Y}$ is

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \frac{1}{N} \sum_{i=1}^{N} [X(i) - \langle \mathbf{X} \rangle][Y(i) - \langle \mathbf{Y} \rangle]. \quad (13)$$

The sample correlation coefficient between $\mathbf{X}$ and $\mathbf{Y}$ is therefore given by

$$\rho(\mathbf{X}, \mathbf{Y}) = \frac{\mathrm{Cov}(\mathbf{X}, \mathbf{Y})}{\delta(\mathbf{X})\, \delta(\mathbf{Y})}. \quad (14)$$

We have $-1 \leqslant \rho(\mathbf{X}, \mathbf{Y}) \leqslant 1$. If it is equal to zero, the random variables $\mathbf{X}$ and $\mathbf{Y}$ are considered uncorrelated. We next define the *correlation distance* by

$$\mathrm{Dist}(\mathbf{X}, \mathbf{Y}) = \frac{1 - \rho(\mathbf{X}, \mathbf{Y})}{2}. \quad (15)$$

Remark 2. We arrange the order of the $F^d(s)$ sequence according to the dictionary order of the 20^K kinds of K strings, then calculate the distance matrix and construct the

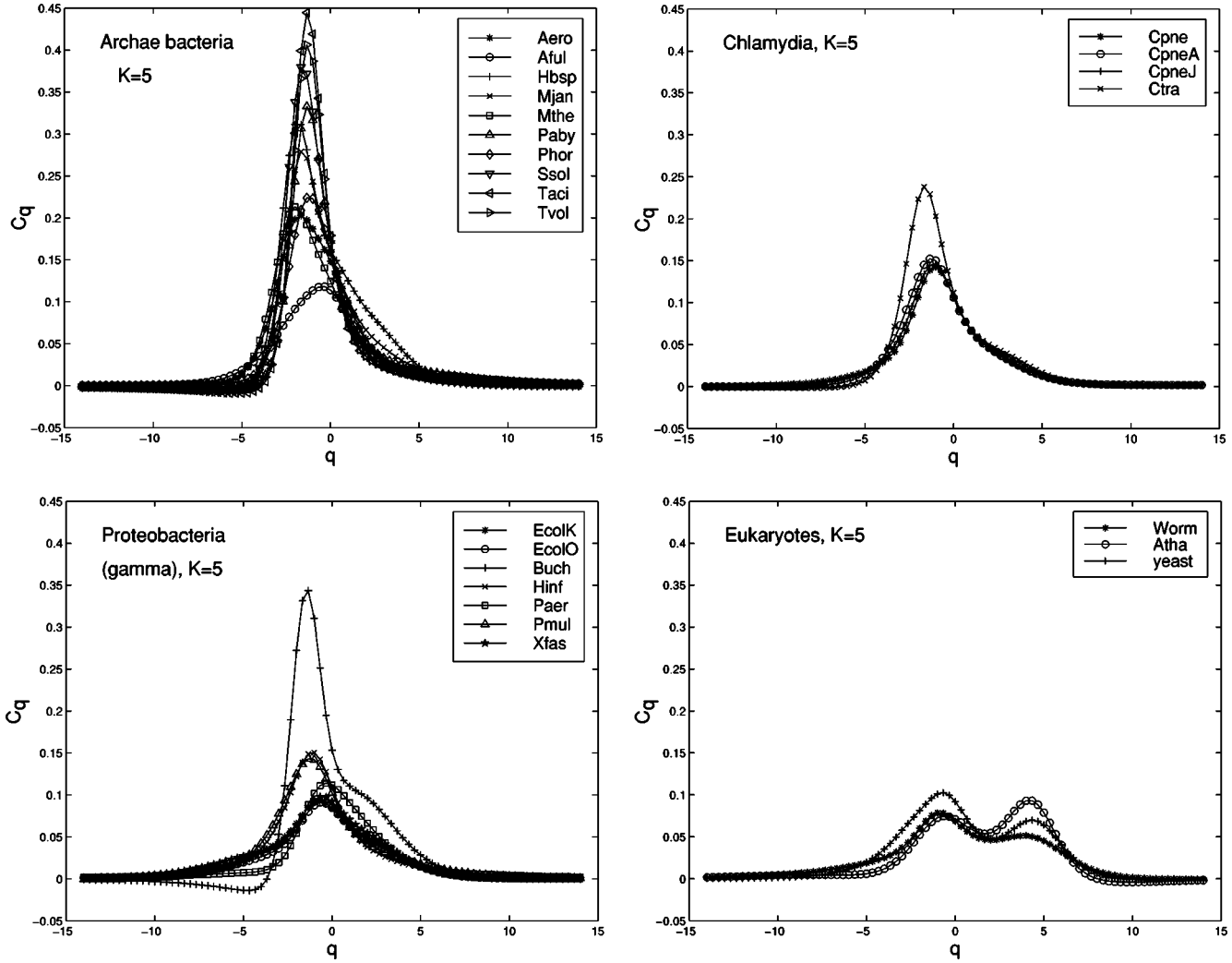

FIG. 5. ''Analogous'' specific heat of measure representation μ of protein sequences of some organisms.

phylogenetic tree. It is easy to see that different orders of the $F^d(s)$ sequence do not change the value of the correlation distance between two organisms using the above definition. A consequence is that different orders of the K strings yield the same phylogenetic tree.

IV. DATA AND RESULTS

Currently there are more than 50 complete genomes of Archaea and Eubacteria available in public databases, for example Genbank at [34]. These include eight Archae Euryarchaeota—*Archaeoglobus fulgidus* DSM4304 (Aful), *Pyrococcus abyssi* (Paby), *Pyrococcus horikoshii* OT3 (Phor), *Methanococcus jannaschii* DSM2661 (Mjan), *Halobacterium* sp. NRC-1 (Hbsp), *Thermoplasma acidophilum* (Taci), *Thermoplasma volcanium* GSS1 (Tvol), and *Methanobacterium thermoautotrophicum* deltaH (Mthe); two Archae Crenarchaeota: *Aeropyrum pernix* (Aero) and *Sulfolobus solfataricus* (Ssol); three Gram-positive Eubacteria (high G+C)—*Mycobacterium tuberculosis* H37Rv (MtubH), *Mycobacterium tuberculosis* CDC1551 (MtubC), and *Mycobacterium leprae* TN (Mlep); twelve Gram-positive Eubacteria (low G+C)—*Mycoplasma pneumoniae* M129 (Mpne), *Mycoplasma genitalium* G37 (Mgen), *Mycoplasma pulmonis* (Mpul), *Ureaplasma urealyticum* (serovar 3)(Uure), *Bacillus subtilis* 168 (Bsub), *Bacillus halodurans* C-125 (Bhal), *Lactococcus lactis* IL 1403 (Llac), *Streptococcus pyogenes* M1 (Spyo), *Streptococcus pneumoniae* (Spne), *Staphylococcus aureus* N315 (SaurN), *Staphylococcus aureus* Mu50 (SaurM), and *Clostridium acetobutylicum* ATCC824 (CaceA). The others are Gram-negative Eubacteria, which consist of two hyperthermophilic bacteria—*Aquifex aeolicus* (Aqua) VF5 and *Thermotoga maritima* MSB8 (Tmar); four Chlamydia—*Chlamydia trachomatis* (serovar D) (Ctra), *Chlamydia pneumoniae* CWL029 (Cpne), *Chlamydia pneumoniae* AR39 (CpneA), and *Chlamydia pneumoniae* J138 (CpneJ); two Cyanobacterium—*Synechocystis* sp. PCC6803 (Syne), and *Nostoc sp. PCC6803* (Nost); two Spirochaete—*Borrelia burgdorferi* B31 (Bbur) and *Treponema pallidum* Nichols (Tpal); and sixteen Proteobacteria. The sixteen Proteobacteria are divided into four subdivisions, which are α subdivision—*Mesorhizobium loti* MAFF303099 (Mlot), *Sinorhizobium meliloti* (smel), *Caulobacter crescentus* (Ccre), and *Rickettsia prowazekii* Madrid (Rpro); β subdivision—

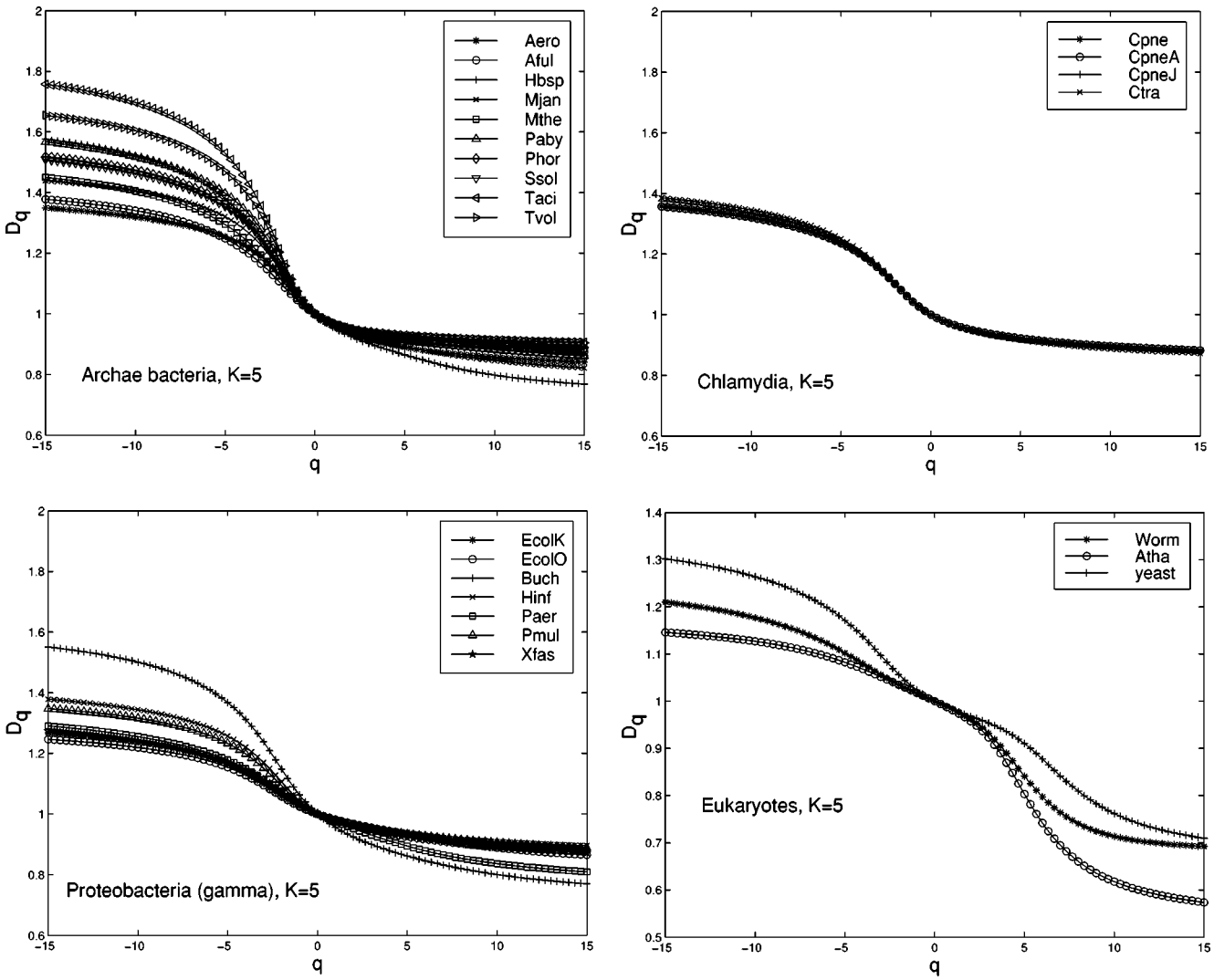

FIG. 6. Dimension spectra of measure μ' (after subtracting some memory) of protein sequences of some organisms.

Neisseria meningitidis MC58 (NmenM) and *Neisseria meningitidis* Z2491 (NmenZ); γ subdivision—*Escherichia coli* K-12 MG1655 (EcolK), *Escherichia coli* O157:H7 EDL933 (EcolO), *Haemophilus influenzae* Rd (Hinf), *Xylella fastidiosa* 9a5c (Xfas), *Pseudomonas aeruginosa* PA01 (Paer), *Pasteurella multocida* PM70 (Pmul), and *Buchnera* sp. APS (Buch); and ϵ subdivision—*Helicobacter pylori* J99 (HpylJ), *Helicobacter pylori* 26695 (Hpyl), and *Campylobacter jejuni* (Cjej). Besides these prokaryotic genomes, the genomes of three eukaryotes: the yeast *Saccharomyces cerevisiae* (yeast), the nematode *Caenorhabdites elegans* (chromosome I-V, X) (Worm), and the flowering plant *Arabidopsis thaliana* (Atha) were also included in our analysis.

We downloaded the protein sequences from the complete genomes of the above organisms and calculated the dimension spectra and analogous specific heat of the measure representations μ and μ' after subtracting some memory. The numerical results showed that it is appropriate to use the measures of $K=5$ (see Ref. [20]). The case $K=6$ is worth trying but beyond our computing power for the time being. For $K=5$, we calculated the dimension spectra, analogous specific heat of μ and μ', and the correlation distances based on $\{F(s)\}$, $\{F^d(s)\}$, and $\{F^{ad}(s)\}$ of all the above organisms. As an illustration, we plot the D_q curves of the measure μ in Fig. 4; and the C_q curves of measure μ in Fig. 5. Because all the D_q are equal to 1 for completely random sequences, it is apparent from these plots that the D_q and C_q curves are nonlinear and significantly different from those of completely random sequences. Hence, all protein sequences from the complete genomes studied are not completely random sequences. We plot the D_q curves of the measure μ' in Fig. 6 and the C_q curves of the measure μ' in Fig. 7.

From the plot of D_q, the dimension spectra of the measures μ and μ' are seen to exhibit a multifractal-like form.

If only a few organisms are considered at a time, we can use the D_q curve to distinguish them. This strategy is clearly not efficient when a large number of organisms are to be distinguished. For this purpose, we find that it is more precise to use C_0, C_1, C_2, in conjunction with the two-dimensional vectors (C_0, C_1) and (C_1, C_2). The distributions of the two-dimensional vectors (C_0, C_1) and (C_1, C_2)

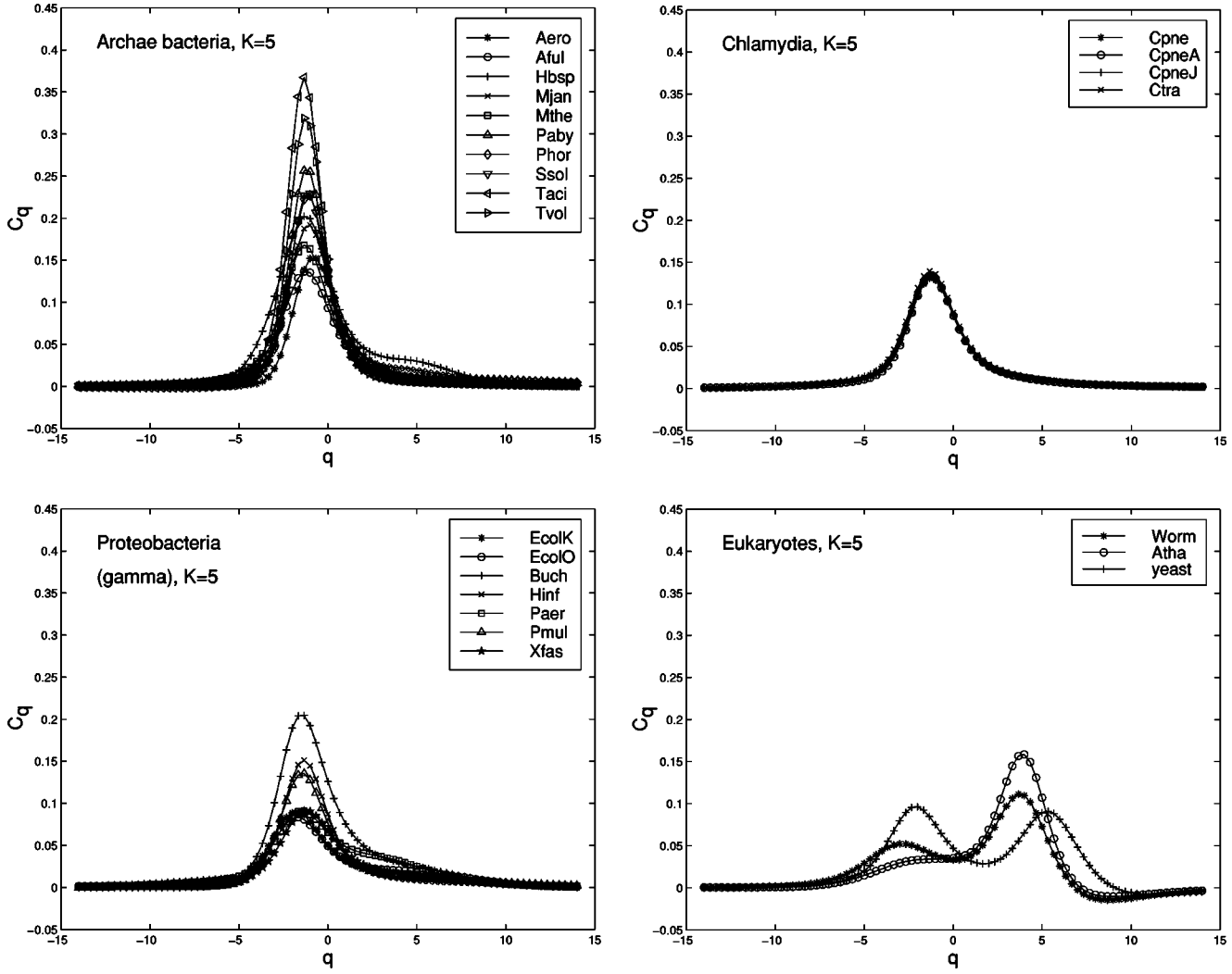

FIG. 7. "Analogous" specific heat of measure μ' (after subtracting some memory) of protein sequences of some organisms.

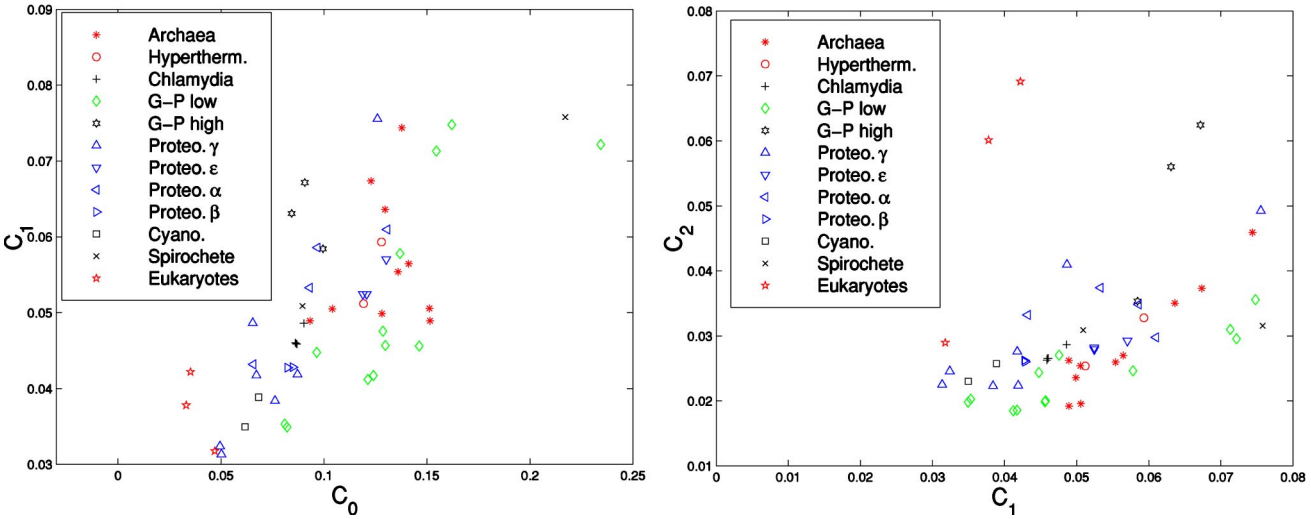

FIG. 8. Distribution of two-dimensional points (C_0, C_1) and (C_1, C_2) of organisms selected.

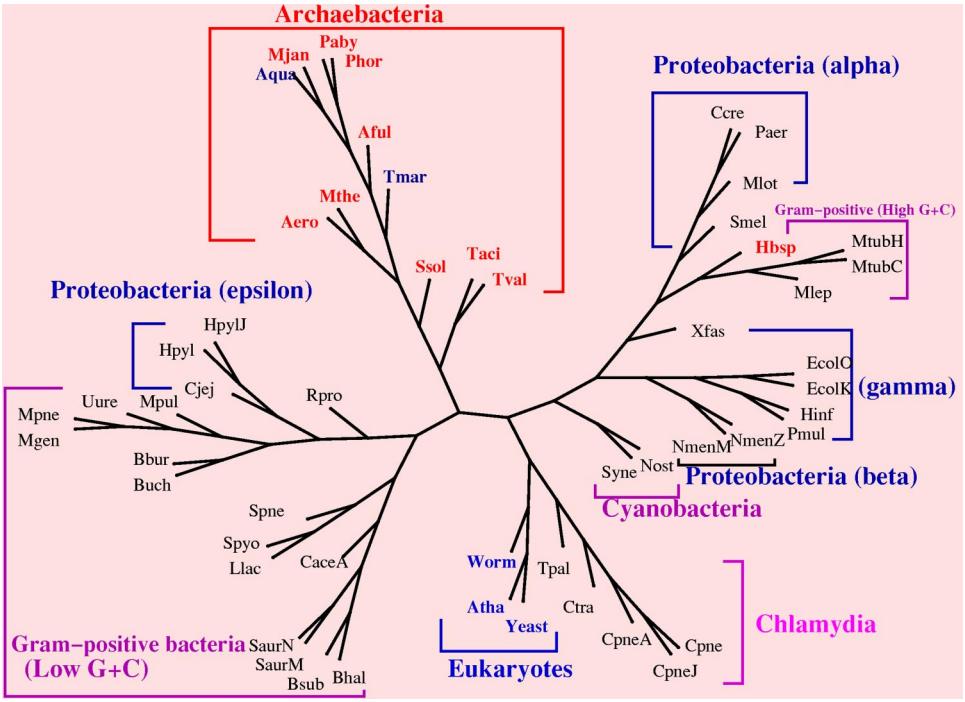

FIG. 9. The neighbor-joining phylogenetic tree based on the correlation distance using $\{F^d(s)\}$ with $K=5$.

based on the measure μ' give more useful patterns for the classification than those based on the measure representation μ. We show the result based on the measure μ' in Fig. 8.

But the above results based on multifractal analysis still do not yield a satisfactory phylogenetic relationship for the organisms selected. For a further improvement, we use the distance matrices from the correlation analysis to construct the phylogenetic tree with the help of neighbor-joining program in the PHYLIP package of Felsenstein [27]. We find that the phylogenetic tree based on the correlation distance using $\{F^d(s)\}$ is more precise than the trees using $\{F(s)\}$ and $\{F^{ad}(s)\}$. We show the phylogenetic tree using $\{F^d(s)\}$ with $K=5$ in Fig. 9.

V. DISCUSSION AND CONCLUSIONS

Deviation of protein sequences from pure randomness or correlation between monomers along the sequences might be of importance [28]. The measure representation of protein sequences provides a simple yet useful vizsualization method to amplify the difference between a protein sequence and a completely random sequence as well as to distinguish protein sequences themselves in more details.

From the measure representation and the values of D_q and C_q, it is seen that there is a clear difference between the protein sequences of all organisms considered here and completely random sequences.

We calculated the D_q and C_q values of two kinds of measures μ and μ' for protein sequences from all organisms selected in this paper for $K=5$. We found that the D_q spectra of all organisms are multifractal-like and sufficiently smooth so that the C_q curves can be meaningfully estimated.

With $K=5$, we found that the C_q curves of all bacteria resemble a classical phase transition at a critical point as shown in Figs. 5 and 7. But the analogous phase transitions of higher organisms are different. They exhibit the shape of double-peaked specific heat function which is known to appear in the Hubbard model within the *weak-to-strong* coupling regime [29].

Although the existence of the archaebacterial urkingdom has been accepted by many biologists, the classification of bacteria is still a matter of controversy [30]. The evolutionary relationship of the three primary kingdoms, namely, archeabacteria, eubacteria, and eukaryote, is another crucial problem that remains unresolved [30].

Figure 8 shows some patterns which are useful for the classification problem, namely, the points corresponding to organisms from the same category are located more closely to each other. But multifractal analysis is still not sufficient to give a satisfactory phylogenetic relationship for the organisms selected. The correlation distance based on $\{F^d(s)\}$ after subtracting some memory from the original information gives a more satisfactory phylogenetic tree. Figure 9 shows that all Archaebacteria except *Halobacterium* sp. NRC-1 (Hbsp) stay in a separate branch with the Eubacteria and Eukaryotes. The three Eukaryotes also group in one branch. Almost all other bacteria in different traditional categories stay in the right branch. At a general global level of complete genomes, our result supports the genetic annealing model for the universal ancestor [31]. The two hyperthermophilic bacteria: *Aquifex aeolicus* (Aqua) VF5 and *Thermotoga maritima* MSB8 (Tmar) stay in the Archaebacteria branch. We noticed that these two bacteria, like most Archaebacteria, are hyperthermophilic. It has previously been shown that Aquifex has close relationship with Archaebacteria from the gene comparison of an enzyme needed for the synthesis of the amino acid trytophan [32].

It has been pointed out [20] that the subtraction of random background is an essential step. Our results show that the subtraction of some memory is also an essential step in our

correlation method. The correlation analysis is more precise than the multifractal analysis for the phylogenetic problem. Although the result from the correlation method of Ref. [20] is slightly better than the result from our correlation method [*Halobacterium* sp. NRC-1 (Hbsp) stays with other Archaebacteria in their phylogetic tree], our algorithm seems simpler, faster and more efficient in using computer space. The reason is that Qi *et al.* [20] used the Markov model to subtract the random background. Hence, their algorithm needs to retain all information of K, $(K-1)$, and $(K-2)$ strings. When K is large, considerable computer space is needed to store this information. On the other hand, our method only requires the information of K strings and the frequencies of 20 kinds of amino acids. Similar to the method in Ref. [20], lateral gene transfer [33] might not affect our results since the correlation method does not depend on the selection of a specific gene.

ACKNOWLEDGMENTS

One of the authors, Z.-G.Y. expresses his gratitude to Professor Bai-lin Hao and Dr. Ji Qi of the Institute of Theoretical Physics of the Chinese Academy of Science for useful discussions on the phylogenetic problem and to Professor E. Canessa of ICTP, Italy, for helpful discussions about multifractal analysis. This work was supported by Grant No. 10101022 of the Youth Foundation of the National Natural Science Foundation of China, QUT Postdoctoral Research Support Grant No. 9900658, the Australian Research Council Grant No. A10024117, and the HKRGC Earmark Grant CUHK No. 4215/99P.

[1] C.M. Fraser *et al.*, Science **270**, 397 (1995).
[2] T.A. Brown, *Genetics*, 3rd ed. (Chapman and Hall, London, 1998).
[3] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992); W. Li, T. Marr, and K. Kaneko, Physica D **75**, 392 (1994).
[4] C.K. Peng, S. Buldyrev, A.L. Goldberg, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley, Nature (London) **356**, 168 (1992).
[5] C.A. Chatzidimitriou-Dreismann and D. Larhammar, Nature (London) **361**, 212 (1993).
[6] V.V. Prabhu and J.M. Claverie, Nature (London) **359**, 182 (1992).
[7] C.L. Berthelsen, J.A. Glazier, and M.H. Skolnick, Phys. Rev. A **45**, 8902 (1992).
[8] Maria de Sousa Vieira, Phys. Rev. E **60**, 5932 (1999).
[9] B.L. Hao, H.C. Lee, and S. Y Zhang, Chaos, Solitons Fractals **11**, 825 (2000).
[10] Z.G. Yu, B.L. Hao, H.M. Xie, and G.Y. Chen, Chaos, Solitons Fractals **11**, 2215 (2000).
[11] B.L. Hao, H.M. Xie, Z.G. Yu, and G.Y. Chen, Physica A **288**, 10 (2001).
[12] H.J. Jeffrey, Nucleic Acids Res. **18**, 2163 (1990).
[13] N. Goldman, Nucleic Acids Res. **21**, 2487 (1993).
[14] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).
[15] Z.G. Yu, V.V. Anh, and K.S. Lau, Phys. Rev. E **64**, 031903 (2001).
[16] E. Canessa, J. Phys. A **33**, 3637 (2000).
[17] V.V. Anh, K.S. Lau, and Z.G. Yu, J. Phys. A **34**, 7127 (2001).
[18] V.V. Anh, K.S. Lau, and Z.G. Yu, Phys. Rev. E **66**, 031910 (2002).
[19] Z.G. Yu and P. Jiang, Phys. Lett. A **286**, 34 (2001).
[20] J. Qi, B. Wang, and B. L. Hao, J. Mol. Evol. (to be published).
[21] M. Li *et al.*, Bioinformatics **17**, 149 (2001).
[22] Z.G. Yu, V.V. Anh, and K.S. Lau, Physica A **301**, 351 (2001).
[23] T. Halsy, M. Jensen, L. Kadanoff, I. Procaccia, and B. Schraiman, Phys. Rev. A **33**, 1141 (1986).
[24] D. Katzen and I. Procaccia, Phys. Rev. Lett. **58**, 1169 (1987).
[25] T. Bohr and M. Jensen, Phys. Rev. A **36**, 4904 (1987).
[26] J. Lee and H.E. Stanley, Phys. Rev. Lett. **61**, 2945 (1988).
[27] J. Felsenstein, The Phylip software, http://evolution. genetics. washington. edu/phylip. html
[28] V. Pande, A.Y. Grosberg and T. Tanaka, Proc. Natl. Acad. Sci. U.S.A. **91**, 12972 (1994).
[29] D. Vollhardt, Phys. Rev. Lett. **78**, 1307 (1997).
[30] N. Iwabe *et al.*, Proc. Natl. Acad. Sci. U.S.A. **86**, 9355 (1989).
[31] C.R. Woese, Proc. Natl. Acad. Sci. U.S.A. **95**, 6854 (1998).
[32] E. Pennisi, Science **286**, 672 (1998).
[33] J.G. Lawrence and H. Ochman, Proc. Natl. Acad. Sci. U.S.A. **95**, 9413 (1998).
[34] See at ftp://ncbi.nlm.nih.gov/genbank/genomes